The Feasibility of Creating a Fully Synthetic Decennial Census Microdata File

Rob Creecy

Census Bureau

July 31, 2009

Joint NSF-Census-IRS Workshop on Synthetic Data and Confidentiality Protection

Outline

Summary of Proposal

- Using the Decennial short form variables
  - Age, Race, Sex, Hispanic, Relationship, Tenure
- Release fully synthetic microdata for persons and households
- For Block level geography
- Tabulations of synthetic data must match already published tables (SF1)

Need for Detailed Demographic Data

- Urban Planning
- Transportation Planning
- Demographic Forecasts
- Tabulations for
  - Custom Geography
  - Custom sets of variables
- Using
  - Micro Simulation
  - Agent based modeling

Users try to synthesize Census data

Breaking the Census "Code":
Reconstructing Original Record-Level Data from Summary Tables

Dmitry Messen
Houston-Galveston Area Council

http://www.trb-appcon.org/TRB2009presentations/s9/06_trb_may_2009_dmitry_messen.ppt

Summary File 1 (SF1)

- Primary source of detailed information from the Census
- Summaries of the 100% Census data compiled into 286 tables
  - Block level population tables (171 tables - label 'P')
  - Block level household tables (56 tables - label 'H')
  - Tract level population tables (59 tables - label 'PCT')
- Compiled from edited, disclosure protected microdata file
- Tables can be thought of as marginal counts of the complete, fully crossed higher dimensional table
- SF1 for 2010 will be very similar to 2000 [1]

[1] see `http://www.census.gov/Press-Release/www/2001/sumfile1.html`

Some SF1 Tables

PCT12A. SEX BY AGE (WHITE ALONE) - Tracts

```
Universe: People who are White alone
 Total:
   Male:
     Under 1 year
     1 year
     2 years
     3 years
     .
     .
     .
     99 years
     100 to 104 years
     105 to 109 years
     110 years and over
     Female:
     (Repeat AGE)
```

Some SF1 Tables

P12A. SEX BY AGE (WHITE ALONE) - Blocks

```
Universe: People who are White alone
 Total:
   Male:
     Under 5 years
     ...
     20 years
     21 years
     22 to 24 years
     ...
     60 and 61 years
     62 to 64 years
     ...
     70 to 74 years
     ...
     85 years and over
  Female:
   (Repeat AGE)
```

Some SF1 Tables (continued)

SEX BY AGE Expansion Tables for Tracts

PCT12A. SEX BY AGE (WHITE ALONE)
PCT12B. SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE)
PCT12C. SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE)
PCT12D. SEX BY AGE (ASIAN ALONE)
PCT12E. SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE)
PCT12F. SEX BY AGE (SOME OTHER RACE ALONE)
PCT12G. SEX BY AGE (TWO OR MORE RACES)
PCT12H. SEX BY AGE (HISPANIC OR LATINO)
PCT12I. SEX BY AGE (WHITE ALONE, NOT HISPANIC OR LATINO)
PCT12J. SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE, NOT HISPANIC OR LATINO)
PCT12K. SEX BY AGE (AMERICAN INDIAN, ALASKA NATIVE ALONE, NOT HISPANIC OR LATIN
PCT12L. SEX BY AGE (ASIAN ALONE, NOT HISPANIC OR LATINO)
PCT12M. SEX BY AGE (NATIVE HAWAIIAN, OTHER PACIFIC ISLANDER ALONE, NOT HISPANIC
PCT12N. SEX BY AGE (SOME OTHER RACE ALONE, NOT HISPANIC OR LATINO)
PCT12O. SEX BY AGE (TWO OR MORE RACES, NOT HISPANIC OR LATINO)

Sex by Age by Race tables for Hispanic's can be derived by subtracting tables PCT12I-PCT12O from tables PCT12A-G

Some SF1 Tables (continued)

SEX BY AGE Expansion Tables for Blocks

P12A. SEX BY AGE (WHITE ALONE)
P12B. SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE)
P12C. SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE)
P12D. SEX BY AGE (ASIAN ALONE)
P12E. SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE)
P12F. SEX BY AGE (SOME OTHER RACE ALONE)
P12G. SEX BY AGE (TWO OR MORE RACES)
P12H. SEX BY AGE (HISPANIC OR LATINO)
P12I. SEX BY AGE (WHITE ALONE, NOT HISPANIC OR LATINO)

Sex by Age by Race tables for Hispanic's can **not** be derived by subtraction at the block level.

Some SF1 Tables (continued)

P18. HOUSEHOLD SIZE, HOUSEHOLD TYPE, AND PRESENCE OF OWN CHILDREN

```
Universe: Households
  Total:
    1-person household:
      Male householder
      Female householder
    2-or-more person household:
      Family households:
        Married-couple family:
          With own children under 18 years
          No own children under 18 years
      Other family:
        Male householder, no wife present:
          With own children under 18 years
          No own children under 18 years
        Female householder, no husband present:
          With own children under 18 years
          No own children under 18 years
    Nonfamily households:
      Male householder
      Female householder
```

Some SF1 Tables (continued)

- Other useful SF1 tables
  - P8. HISPANIC OR LATINO BY RACE
  - P26, P26A-P26I. HOUSEHOLD TYPE BY HOUSEHOLD SIZE
  - P27, P27A-P27I. RELATIONSHIP BY HOUSEHOLD TYPE (INCLUDING LIVING ALONE)
- Generally, there is more detail for Tracts than blocks.
- Variables have more detail - race (e.g. multirace, Asian categories), Hispanic origin, Native American Tribe.
- SF1 is a complex, interrelated set of tables - generating good synthetic data requires studying these tables carefully to help ensure consistency.

Technical Approach for Synthesis

- Fit a model for cell proportions, θ, using SF1 tables as marginal constraints.
- Draw a random vector of cell proportions θ^*, from the posterior distribution of the model.
- Draw a random vector of cell counts, x^*, using parameter θ^*.
- Release x_j^* records with variable values corresponding to index j (or just release x^*).
- Repeat for multiple imputations.

Technical Approach for Synthesis

- Fit a model for cell proportions.
- Pick a set of SF1 tables.
- For these tables, identify the full set of p underlying categorical variables, $Y_1, Y_2, \ldots, Y_p$, with Y_k having d_k distinct values. Each SF1 table is a cross-classification of a subset of $Y_1 \ldots Y_p$.
- Let $x = (x_1, x_2, \ldots, x_D)$ be the counts in the contingency table of size $D = d_1 \times d_2 \ldots \times d_p$ defined by the cross-classification of $Y_1, Y_2, \ldots, Y_p$.
- Assume x is multinomial with $x|\theta \sim M(n, \theta)$, and the $\theta = (\theta_1, \theta_2, \ldots, \theta_D)$ are the cell probabilities.

Technical Approach for Synthesis

Fit a loglinear model for the cell probabilities, θ, with design matrix X.

$$\log(\theta) = X\beta$$

- Include the interaction of all variables defining each SF1 table as terms (configurations) in the model matrix X.
- The data for each SF1 table are the sufficient statistics for the model, so no original microdata are needed to fit the model.
- Use Iterative Proportional Fitting (IPF) to get MLE's (or posterior modes if using priors) of the cell proportions, $\hat{\theta}$.

Technical Approach for Synthesis

- Draw a random vector of cell proportions θ^*, from the posterior distribution of the model, which is Dirichlet-like with parameter $n\hat{\theta}$.
  - Either use Bayesian Iterative Proportional Fitting (BIPF), or
  - Rake θ^* back to the margins so that $nX'\theta^* = X'x$.
- Draw a random vector of cell counts, $x^* \sim M(n, \theta^*)$.
  - Rake x^* back to the margins so $X'x^* = X'x$
  - Use an unbiased controlled rounding procedure to obtain integer cell counts.
  - Exact controlled rounding is not always possible - we have developed some new unbiased approximate methods.

Example - Block, Sex, Age, Race, Hispanic

Pick a tract and all the blocks within that tract, (3922 persons in 38 blocks)

- Use SF1 tables
  - PCT12A-PCT12O (Sex(2), Age(23), Race(7), Hispanic(2))
  - P12A-P12G (Block(38), Sex(2), Age(23), Race(7))
  - P12H (Block(38), Sex(2), Age(23), Hispanic(2))
- The complete table x is of dimension $D = (38 \times 2 \times 23 \times 7 \times 2) = 24472$
  - The table is very sparse, just 1980 non-zero cells
- Fit a loglinear model using IPF that includes terms ((Sex, Age, Race, Hispanic), (Block, Sex, Age, Race), (Block, Sex, Age, Hispanic))

Example - Block, Sex, Age, Race, Hispanic Results

- Comparison of fitted table to original table
  - All original cell counts are within 2.6 of the fitted values.
  - Almost all of the original non-zero cell counts (88.6%) are exactly equal to the rounded, fitted cell counts.
  - The fitted table has 436 more cells with positive probability than the original, with a mean of .33 and a max of 1.71 .

Difference	0	1	2	3
Num. of cells	1755	212	11	2
Pct. of cells	88.6	10.7	0.6	0.1

Table: Absolute differences between original cell counts and rounded, fitted cell counts

Example - Block, Sex, Age, Race, Hispanic Results

- Comparison of randomized tables to original table
  - Simulation of random synthetic procedure, 100 repetitions
  - As expected, the mean of the simulated cell proportions equals the fitted cell proportions and the mean of the simulated cell counts equals the fitted cell means.
  - All original cell counts are within 4 of the simulated cell counts.
  - On average, (84.54%) of the original cell counts are exactly equal to the simulated cell counts.

Difference	0	1	2	3	4
Num. of cells	1673.94	267.59	32.61	5.24	0.62
Pct. of cells	84.54	13.51	1.65	0.26	0.03

Table: Absolute differences between original cell counts and randomized cell counts, averaged over 100 repetitions.

Improving accuracy of synthetic data

- Using just published margins in the model for synthetic data leads to some bias in the internal cells of the complete table.
- The amount of bias and variance for the proposed procedure needs to be assessed.
- Using (Dirichlet) priors based upon the original data will reduce bias, at the possible cost of increasing disclosure risk.

Synthesizing Households

- Bottom Up
  - Synthesize persons first
  - Model and synthesize household types, defined by
    - number of persons
    - married or not married
    - family or not family
    - with or without children
  - Model the relationship between household types and person characteristics
- Top Down
  - Synthesize household types first
  - Synthesize persons conditional on household type
  - Issue: how to make marginal tables of synthesized persons consistent with SF1

Conclusion

- Creating a reasonably accurate synthetic person microdata file using just the published tables as margins is feasibile.
- More research could be done on the disclosure risk of the current swapping method given the possibility of users accurately modeling the unpublished cells.
- Improving the accuracy of the synthetic data by using priors derived from the original microdata should be possible, but more research is needed on the best technical approach and how to best trade off disclosure risk and analytic validity.
- Creating synthetic households that are consistent with the synthetic person microdata is a more difficult problem because of the much higher dimensionality of the problem, but several approaches seem promising.